DOCUMENT RESUME

ED 067 143 LI 003 891

AUTHOR	Williams, Martha E.
TITLE	Handling of Varied Data Bases in an Information Center Environment.
INSTITUTION	Illinois Inst. of Tech., Chicago. Research Inst.
PUB DATE	23 Jul 71
NOTE	24p.; (0 References); Presented at the Conference on Computers in Chemical Education and Research, Northern Illinois Univ., DeKalb, Illinois
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Computer Programs; *Data Bases; *Information Centers; Information Processing; *Information Services; Search Strategies
IDENTIFIERS	*Key Letter In Context Index; KLIC Index; Selective Dissemination of Information

ABSTRACT

Information centers exist to provide information from machine-readable data bases to users in industry, universities and other organizations. The computer Search Center of the IIT Research Institute was designed with a number of variables and uncertainties before it. In this paper, the author discusses how the Center was designed to enable it to accommodate the many variables it would face in providing different services to diverse users. The system design is discussed in terms of the unpredictable future and in terms of the users to be served. User aids that were developed (search manual, truncation guide, frequency lists, KLIC index and bigram frequency list), and communication with the user are discussed. The reasons for using a selective dissemination of information service (SDI) are presented. (Author/SJ)

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED 067143

HANDLING OF VARIED DATA BASES IN AN INFORMATION CENTER ENVIRONMENT

by

Martha E. Williams
IIT Research Institute
10 West 35th Street
Chicago, Illinois 60616

Presented at the

Conference on Computers in
Chemical Education and Research
Northern Illinois University
DeKalb, Illinois

July 23, 1971

LI 003 891

HANDLING OF VARIED DATA BASES
IN AN INFORMATION CENTER
ENVIRONMENT

DESIGN FOR UNPREDICTABLE FUTURE 1

File Structure 1
Preprocessors 4
Hardware 5
Software Modularity 5
Machine and Installation Independence 5

USER ORIENTED DESIGN 7

Data Base Options 7
Profile Options 7
Profile Features 7

Truncation 7
Logic 9
Linking or Grouping of Terms 9

USER AIDS 9

Search Manual 10
Truncation Guide 10
Frequency Lists 10
KLIC Index 12
Bigram Frequency List 12

COMMUNICATION WITH USER 14

REASONS FOR USING THE SDI SERVICES OF AN INFORMATION CENTER 15

Coverage 15
Thoroughness of Search 15
Consistency of Search 16
Interdisciplinariness 16
High Recall 16
Cost-Effectiveness 17
Speed and Regularity 18
Timeliness 18
Multiplicity of Data Bases 18
Automatic Preparation of Files in Standardized Format 18
Cost of Data Base Preparation and Operation of an SDI System vs. Subscriptions 19

Cost of Data Base Preparation 19
Cost of Operation of an SDI System 19
Cost of Subscription 21

DESIGN FOR UNPREDICTABLE FUTURE

Information centers exist to provide information from machine-readable data bases to users in industry, universities and other organizations. The Computer Search Center of IIT Research Institute was designed with a number of variables and uncertainties lying before us. We were not able to predict the data bases to be used, the character sets, the character codes, data base tape formats, data base record formats, data elements, and data base content. Nor could we predict the hardware we might use, user groups, user requirements, vocabulary, search requirements, output sorts, output formats, and output media. These are a few of the variables we saw. Our design had to accommodate all of them. We did not pre-determine which data bases and what user groups we would provide services to. We wanted to be in a position to handle any data base for which there is a significant market.

As we all know, search programs for handling machine-readable data bases are expensive to develop and expensive to maintain. We had no desire to incur the expense of maintaining multiple search programs. For this reason we developed a general purpose search program that would handle virtually any of the machine readable data bases containing natural language information.

When handling multiple data bases one is very likely to encounter multiple character sets and multiple character codes. The tape formats and record formats differ from data base to data base. In fact, they differ within data bases that are produced by the same organization. The data elements contained on the tapes vary considerably from tape to tape.

File Structure

In order to handle multiple data bases we had to determine a file structure that would accommodate all of the variables. The IITRI standard format does precisely that. With this format each record is composed of a key, directory, and character string. The key contains the volume, issue, and citation number as given by the data base supplier, and the directory identifies each type of element contained in the record, according to IITRI data type codes. Some of the data type elements are listed.

DATA TYPE CODES	DATA ELEMENT
01	SOURCE INFORMATION --CODEN --JOURNAL REFERENCE --PAGINATION --DATES
02	TITLE OF ARTICLE
03	AUTHOR(S)
04	SHORT JOURNAL TITLE
05	KEYWORD(S) INDEX TERMS CA SECTION NUMBER
06	CA REGISTRY NUMBER
07	MOLECULAR FORMULA
08	CORPORATE AUTHOR
09	ABSTRACT TEXT
10	BA CROSS CODE
11	BA BIOSYSTEMATIC INDEX
12	EI CARD-A-LERT CODE
13	ORIGINAL LANGUAGE AVAILABILITY PUBLISHER PRICE DATE OF MEETING PARENT JOURNAL ORIGINAL ABSTRACT SOURCE

Figure 1. Data Elements and IITRI Data Type Codes

Key: 7416-081368 (Volume, Issue and Abstract Number)

Directory:			
1	1	26	(CODEN)
4	27	14	(Journal)
2	41	76	(Author(s))
3	117	60	(Title)
8	177	51	(Corp. Author)
5	228	40	(Index Terms)
13	268	17	(Language)

String:

JPCHAX/75/3/325-30/000071/J. PHYS. CHEM.VIBRONIC EFFECTS IN THE INFRARED SPECTRUM OF THE ANION OF TETRACYANOETHYLENEDEVLIN, J. PAUL$MOORE, JESSE C.$SMITH, DONALD$YOUHNE, YOUNG$DEP. CHEM., OKLAHOMA STATE UNIV., STILLWATER, OKLA.$CA073000$ IR SPECTRA ALKALI METAL SALTSORIG. LANG.: ENG

Complete Record Appears on Tape as:

7416-081368 1 1 26 4 27 14 2 41 76 3 117 60 8 177 51 5 228 40 13 268 17 JPCHAX/75/3/325-30/000071/J. PHYS. CHEM. VIBRONIC EFFECTS IN THE INFRARED SPECTRUM OF THE ANION OF TETRACYANOETHYLENEDEVLIN, J. PAUL$MOORE, JESSE C.$SMITH, DONALD$YOUHNE, YOUNG$DEP. CHEM., OKLAHOMA STATE UNIV., STILLWATER, OKLA.$CA073000$ IR SPECTRA ALKALI METAL SALTSORIG. LANG.: ENG

Figure 2. IITRI Formatted Citation

In the directory the code is followed by the starting position for the actual data, and an indication of the number of characters required by the data. Thus, in Figure 2, for the record having Citation Number 81368 of Volume 74, Issue 16, in Chemical Abstracts Condensates there is a CODEN which starts in position 1 and is 26 characters long. The next kind of data element included in the record is a Journal Name which is a type 4 code. The actual data starts in position 27 which is one position beyond the end of the CODEN data and is 14 characters long. The next data element is the Author Name, which is a type 2 data code, and starts in position 41 which is one position beyond the end of the journal data. Author data is 76 characters long, and on down the line. If you follow through Figure 2, the format becomes obvious. The string portion of Figure 2 shows how the actual data for this particular reference is contained in IITRI format on tape, and the complete record which appears in the lower portion of Figure 2 shows the entire key directory and character string for the particular record as it appears on tape.

The use of data element codes allows us to handle multiple varied data elements. The system also allows us to add new data elements and new data type codes as they arise. We have no way of knowing what new data elements suppliers may include in their tapes a few years from now. However, we have allowed for 9,999 different data type codes. It is unlikely that we will be unable to accomodate any new data element that should come into existence.

The standard IITRI format is employed for any data base processed. Our method for handling multiple data bases is to write a pre-processor program for each different data base that is handled in the system. The pre-processor program reformats the data that is contained on the supplier data base and puts it into IITRI format. In that way every data base looks the same to the search program, and all data bases can be handled by one and the same search program.

Preprocessors

The cost of writing pre-processor programs is very slight compared to the cost of maintaining multiple search programs. A preprocessor or format conversion program takes about 2 man weeks of programming effort to write. We initiated this system of format conversion for a generalized search program three years ago. It has worked very well. It has allowed us to accommodate changes in data bases whenever the need arose. Fortunately or unfortunately the data base suppliers do make frequent charges in the format and content of their data bases. These changes may not be large or terribly significant, however, as far as computer programs are concerned, any change can be

disastrous if one is not able to accommodate it.

The use of preprocessor together with a flexible and expandable file structure, have made it possible to accommodate multiple data bases with varying contents, varying data elements, differing character sets and character codes, and different tape formats and record formats.

Hardware

The next area of unpredictability that affected our design is related to hardware. We, like many organizations, have changed computers with some regularity. We did not want to be locked to any particular configuration, nor to develop our software in an assembly language that would be very machine and installation oriented. The system had to be as machine independent and installation independent as possible. For this reason we decided to adopt a higher level compiler language and to use a family of computers that enjoyed wide acceptance throughout the country and internationally--IBM 360. The compiler language we adopted was PL/1, a language that is highly amenable to handling natural language data. PL/1 compilers are available on IBM 360 series computers and they have been announced for Burroughs and CDC hardware.

Software Modularity

We wrote the programs in a modular fashion in order to easily accommodate changes to any portion of the system. The principal modules in the systems include: (1) format conversion for reformatting data bases; (2) a profile editing module, that handles the terms, logic and user related information that are associated with each profile; (3) a search module which is the heart of the system and matches user profile terms against terms on the data base within the requirements of the profile; (4) a formatting module for preparing output in dissemination format; (5) a module that sorts and prints the output; and (6) a statistics gathering module.

Machine and Installation Independence

The reasons for developing our software in a machine independent and installation independent fashion were that we anticipated our own hardware might change--which it did--, we wanted to be able to install our system at other organizations that had a need for running SDI within their own facilities, and we wanted to be able to provide profile-writing training courses and workshops at other locations. Successful achievement of machine and installation independence with IBM equipment is evident from the fact that we have run our programs at nine different computer facilities with different: hardware,

computer models, versions of the operating system, peripherals, and releases of the PL/1 compiler. Figure 3 indicates the varieties we encountered in the nine facilities. In no case did we have any real difficulty. Preparation of appropriate JCL is usually all that is required.

Hardware:	IBM 360	Models:	40 50 65 67 75
	IBM 370	Model:	155
	Any computer with PL/1 Compiler		
Processors:			MFT MVT PCP HASP
Operating System Versions:			15-16 17 18 19 19.6 20
PL/1 Compiler Releases:			4.1 5 5.2

Figure 3. Machine and Installation Independence of IITRI Software

USER ORIENTED DESIGN

Data Base Options

The current awareness or SDI (Selective Dissemination of Information) system has been operational since the spring of 1969. Data bases handled by the Computer Search Center (CSC) are Chemical Abstracts Condensates, Biological Abstracts, BioResearch Index, and Engineering Index's COMPENDEX. Two more data bases are planned for the near future and other data bases will be added depending on user needs. We will handle any data base for which there are enough subscribers. Searches are conducted and output sent to users weekly, biweekly or monthly in accordance with the frequency of the particular data base to be searched.

Profile Options

One may include searchable elements as positive or negative search terms, i.e., one may require the presence or absence of any particular search term to qualify a citation as a "hit" citation. Among the searchable elements are those shown in Figure 1. The search terms may be single words, multi-word terms, phrases or portions of words, i.e., any legitimate character string. The range of options available to users for their profiles is shown in Figure 4.

Profile Features

The principal features built into the system to achieve effective profiles are the following: All forms of term truncation; full free form Boolean logic with any degree of nesting; grouping or linking of similar terms; and weighting of terms according to user assignment of relevance. Weighting is used to indicate relative importance of terms, separate closely related concepts and to sort output in relevance order.

Truncation

The terms themselves may be truncated to facilitate retrieval of terms containing common fragments. The four truncation modes allowed are shown in Figure 5. The usefulness of right truncation is usually readily understood. Right truncation is used to select singular, plural and other forms of words that contain a common stem. The usefulness of left truncation is not so obvious but it can be readily demonstrated. For example one might use the left truncated term *MYCIN to represent antibiotics and retrieve:

Terms - anything other than single character

- Single word
- Multi word
- Phrase
- Fraction of term
- Symbol or acronym

Kinds of Terms - anything on the data base

- Author
- Company Name
- Country
- Language
- Date
- CODEN
- BA Cross Code
- BA Biosystematic Index
- Keywords/Index terms
- CA section number
- CA registry number (when available)
- Molecular formula (when available)
- EI Card-a-lert Code
- Abstract or Text term
- Any term on any data base searched

Sorting of Output

- Author
- Weight
- Citation number

Output Media

- 5" x 8" cards
- Paper
- Multilith masters
- Tape
- Microfilm

Figure 4. Profile Options

Mode	Function	Example
none	requires exact match of a term	term AZO
left	allows substitution of any prefix on the term	* term DI AZO
right	allows substitution of any suffix on the term	term * AZO XY
both	allows substitution of any prefix and/or suffix	* term * DI AZO METHANE

NOTE: * denotes truncation

Figure 5. Truncation Modes

ACTOMYCIN
ANTIMYCIN
BIOMYCIN
ERYTHROMYCIN
NEOMYCIN
STAPHYLOMYCIN
STREPTOMYCIN

The one search term *MYCIN substitutes for 20 to 30 specific terms which are not shown. The use of simultaneous left and right truncation would pick up all of the above terms plus plural forms e.g., 'STREPTOMYCINS', and those with terminal punctuation e.g.,"ACTINMYCIN.".

Logic

Profile terms are related to one another by means of logic symbols. The logic symbols used for the logical operators AND, OR, and NOT are:

Logical Operators	Symbol
AND	&
OR	|
NOT	¬

The logic expressions for profiles can be as specific and involved as is necessary to express the user's question. While most expressions are relatively simple, any expression can be handled by the system. For example, the following expression would be legitimate:

(((A&B) | (C|D|E|F)) &¬G) | ((H&I) &¬J)

Linking or Grouping of Terms

In order to simplify the writing of a profile, similar terms may be linked together by a link code and then the group will be referred to by the single letter code. An example is given in Figure 6, where a user has requested information on reactions of halogens and alkali metals.

User Aids

In addition to providing many options for preparing profiles we have developed a number of aids to help users write profiles, select words, truncate words, and word fractions. Index Terms and Hit Terms are printed on each output card to provide the user with information for revising his profile.

Terms	Link Code		Terms	Link Code
Halogen	A		Alkali metals	B
Halide	A		Lithium	B
Fluorine	A	AND	Sodium	B
Chlorine	A		Cesium	B
Bromine	A		Potassium	B
Iodine	A		Rubidium	B

Rather than:

(Halogen | Halide | Fluorine | Chlorine | Bromine | Iodine) & (Alkali metals | Lithium | Sodium | Cesium | Potassium | rubidium)

User specifies: (A&B)

Figure 6 - Linking or Grouping of Terms in a Profile

Search Manual

The CSC Search Manual explains the basic techniques of profile writing, including: question formulation, concept identification, concept expansion, profile refinement and profile modification. Methods for using search terms, truncation, logic, links and weighting are described. A Supplemental Guide has been written for each data base. The guide demonstrates profile writing tailored to the specific data base.

Truncation Guide

A Truncation Guide illustrates where to truncate a term in order to retrieve the maximum relevant words with the minimum noise. Figure 7 from the profile Truncation Guide demonstrates the retrieval ability of various forms of terms related to the concept ANALYSIS.

Frequency Lists

A Frequency List in frequency order and a Frequency List in alphabetic order have been prepared for each data base. They are used to assist in selection of search terms. A high frequency term will produce a high volume of hits unless it is combined with another search term or assigned a low weight. A low frequency word might be used independently. Frequency lists are used as rough indicators of the volume of output one might expect to receive for specific terms. Our Frequency Lists have been prepared for one volume of each data base.

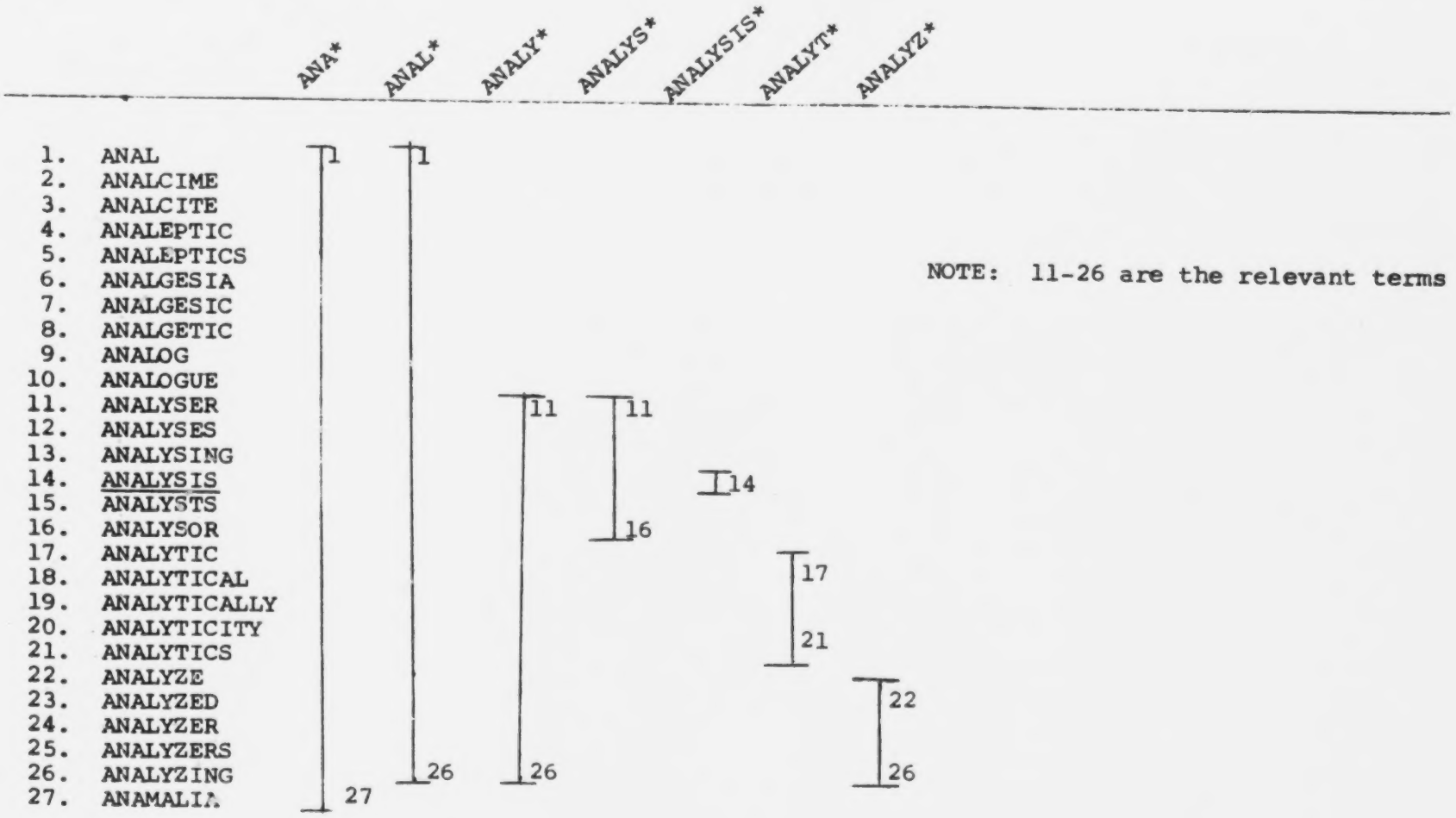

Figure 7. Truncation Guide Entries For The Concept ANALYSIS

KLIC Index

A KLIC Index (Key Letter-In-Context Index) has been prepared for each data base. The KLIC Index is a lexicographic ordering of all terms in a data base by each character (alpha, numeric or special) in the term or character string. It is a permuted arrangement sorted by character with the balance of the term wrapped around it. See Figure 8. The KLIC is used for linguistic research and as a user aid. By consulting the KLIC one can determine the retrieval capability of a particular letter combination or term fragment. The KLIC is used to identify letter combinations that are highly specific and would therefore be discriminating search terms e.g., the character string *YBD* does not occur anywhere in the CA or BA data bases except in the term MOLYBDENUM (Note: in a literary data base it would occur in the mythological character SCYLLA and CHARYBDIS). Thus, *YBD* could be used as a search term for molybdenum. On the other hand, letter combinations that occur frequently in many irrelevant terms should be avoided e.g., the letters RNA for ribonucleic acid could be used as a search term assuming one did not specify simultaneous left and right truncation *RNA*. The simultaneous truncation mode would retrieve more than 200 irrelevant words. Some of these are:

ALTERNATE

BARNACLE

CARNATION

DIURNAL

FINGERNAIL

MATERNAL

Bigram Frequency List

The KLIC Index indicates where letter combinations occur and our Bigram Frequency List, which provides a frequency count for every two letter combination in the data base, indicates how often each bigram occurs.

FILMED FROM BEST AVAILABLE COPY

KEY-LETTER-IN-CONTEXT INDEX

ANIM ALS' /	ASCH ALT /
CRYST ALS' /	DICOB ALT /
MINER ALS' /	ROCKS ALT /
ORBIT ALS' /	HIGH-S ALT /
CHEMIC ALS' /	SULFOS ALT /
MATERI ALS' /	TRICOB ALT /
TION-RADIC ALS' /CA	ACYLCOB ALT /
ALSACE /	FUSED-S ALT /
/ ALSACIENNES	WATER-S ALT /
B ALSAM /	IRON-COB ALT /
B ALSAMINA /	MOLTEN-S ALT /
V ALSARIA /	RADIOCOB ALT /
S ALSBURY /	ZINC-COB ALT /
E ALSE /	AMMINECOB ALT /
CAT ALSE /	AQUEOUS-S ALT /
W ALSER, /	ORGANOCOB ALT /
S ALSES-LA /	SULFATE-S ALT /
W ALSH /	COPPER-COB ALT /
W ALSH, /	IRON(III)-S ALT /
TH ALSIMINE /	NICKEL-COB ALT /
DIHYDROTH ALSIMINE /	AMINATOCOB ALT /AL
ALSINOIDES /	ARBIDE-COB ALT /C
H2O / NA ALSIO4-SIO2-	NICKEL-COB ALT /IRON-
ALSIO6 /	YRAZINECOB ALT /P
PERIST ALSIS /	LADIUM-COB ALT /PAL
LI ALSI2O6 /	ANGANATE/S ALT /PERM
CA ALSI2O8 /	ATINUM-COB ALT /PL
NA ALSI3O8 /	CHEIDEANST ALT /S
CZOCHR ALSKI /	MARIUM-COB ALT /SA
LER-NAPIER ALSKI /BISCH	RIALLYLCOB ALT /T
MICH ALSKI, /	TTRIUM-COB ALT /Y
TEMP ALSKI, /	S ALT. /
KATCH ALSKI, /	/ COB ALT(III)UREA
CZOCHR ALSKI'S /	YANIDE)COB ALT(I) /ISOC
P ALSKY, /	LT(III)/COB ALT(II /COBA
W ALSMANN, /	COB ALT(II) /
C ALSNS /	OXIDE-COB ALT(II) /
ALSO /	YRAZINECOB ALT(II) /P
S ALSOLA /	LIGANDCOB ALT(II). /
S ALSOLINE /	PTIDE /COB ALT(II)-DIPE
ALSON /	AMIC / COB ALT(II)-GLUT
INGM ALSON, /	ILOTR /COB ALT(II)-NITR
H ALSTEDII /	,4'', /COB ALT(II)-4,4'
ALSTEN, /	LT(II /COB ALT(II)/COBA
/ ALSTONERINE,	COB ALT(II), /
/ ALSTONERINES	AMMINECOB ALT(III /HEX
ALSTONIA /	TAMMINECOB ALT(III /PEN
R ALSTONITE /	COB ALT(III) /
ES / ALSTOPHYLLIN	AQUOCOB ALT(III) /
G ALSTYAN, /	COB ALT(III). /
V ALSUGANA /	ICALL /COB ALT(III)-OPT
NAPHTH ALSULFONIC /	MOTED /COB ALT(III)-PRO
M ALT /	COB ALT(III), /
S ALT /	S ALT) /
BAS ALT /	COB ALT) /
COB ALT /	ANST ALT) /

Figure 8. KLIC Index Entries

Communication with User

In order to maintain good rapport with users and to be sure that their profiles are functioning efficiently to provide the desired information, CSC uses many avenues of communication with users. Among them are:

- Unlimited profile changes
- Low cost profile switch
- Evaluation reports
  (95% response implies use)
- Feedback cards
- Continuous precision calculations
- Telephone contact
- Comments on profiles to suggest changes in logic weighting, and grouping of terms, or to suggest use of new data elements or new terminology

The concern for users is of extreme importance to information centers. Information systems are designed to be used and if the clients are not satisfied with the service they will not use it.

REASONS FOR USING THE SDI SERVICES OF AN INFORMATION CENTER

There are some scientists that have no need for or cannot use SDI. There are those who have no need to monitor or use the new developments in a field, those for whom there is no appropriate data base, those whose area of interest is so extremely narrow and highly specialized that the appropriate terminology is unlikely to be used by authors in titles or by indexers in index terms, and those who can truly monitor the publications in a field by scanning a handleable number of journals. The members of this group are not nearly so numerous as they might think.

It remains, then that there are many scientists who do have a need for SDI, and there are advantages and reasons for using the SDI services of an information center. They are: (1) the extensiveness and inclusiveness of the broad coverage, (2) thoroughness of search; (3) interdisciplinariness; (4) high recall; (5) cost effectiveness; (6) speed and regularity; (7) timeliness; (8) multiplicity of data bases; (9) automatic preparation of files in standardized format; and (10) cost of data base preparation and operation of an SDI system vs. subscriptions.

Coverage

The first and most obvious reason for SDI, is that for the most part, it simply cannot be done manually any more because the volume of publications is so large. While formerly one could "eyeball" Chemical Abstracts or other secondary sources to cover the literature in his area of specialization, this would now be a monumental task. The journal coverage of data bases such as CAS Condensates, BA Previews and others is quite extensive; for example, in the field of chemistry a year's worth of Condensates includes 300,000 chemical references taken from approximately 20,000 journals -- such extensive coverage cannot be duplicated elsewhere. In the field of biology, BA Previews includes approximately 250,000 references selected from approximately 8,000 journals. In engineering the COMPENDEX data base produced by Engineering Index includes 75,000 references a year from 3,500 journals. There are many additional data bases providing extensive coverage in their respective fields.

Thoroughness of Search

Machine searching is more thorough than human searching. The computer will search every profile term against every citation on the data base - no reference is overlooked and human fatigue is not a factor. Once the term matches have been made then the "hit" citations are checked to see if they satisfy the logic the user employed to relate the terms to each other.

Consistency of Search

A machine search evaluates every citation in the data base in exactly the same manner. The search strategy and criteria for selection are employed in the same way for the first citation on a tape and for the last citation. While a human searcher is likely to be affected by fatigue and boredom, the computer is not.

Interdisciplinariness

One of the most important reasons for using a machine search is related to the interdisciplinary character of research. The user may be working in a field that requires coverage of several subfields within a data base or several data bases. For this reason it would be very difficult to identify and manually search all the appropriate portions of a data base. It is also becoming increasingly difficult to anticipate which area of subspecialization a particular journal article might be assigned to within a secondary source. For example, an article dealing with a particular air pollutant might find its way into any one of a number of sections in CA -- all of which would be correct.

An organic chemical could be assigned to one of the organic chemistry sections; if detected by some analytical device it could be attributed to one of the analytical chemistry sections; as an air pollutant to the air pollution section; if it were inhaled and produced a biologic effect it might be assigned to toxicology; if it were deposited on a body of water it would apply to water pollution; and if deposited on the ground it might apply to the section on soil and plant growth.

This is just one example but there are many and this interdisciplinary factor has been cited by our users repeatedly as being a real advantage. Many users, who had manually searched CA for years, thought they were doing a good job, and thought they knew exactly which sections of CA would be appropriate, were very surprised to find that the SDI search located relevant references from sections they would never have examined.

High Recall

Because of the previously mentioned advantages it is possible for the user to achieve a higher recall with SDI than he could manually. Naturally, the value of high recall vs. high precision varies from user to user depending on his objective in using the service. The user or the profile coordinator acting for the user, must weigh the tradeoffs and determine whether he can afford to miss a few relevant references in order to achieve high precision or whether he can afford to retrieve

some irrelevant references in order to ensure his not missing anything.

An example of the recall ability of the system is the case of one of our user companies that maintained a manual search system parallel to the SDI we provided and compared the output for over a year. For many years and throughout the experimental year 20 bench chemists divided up the sections of CA and searched for references relevant to the companies areas of research. The output of their search was forwarded to their technical library as was the output of our SDI. Eleven profiles were written and run against Condensates. The results of the study was that the manual search identified 5% more relevant references than the machine and the machine identified 15% the chemists missed. Simply, if total relevant references identified by both sources is considered 100% then the recall for the manual search by professional chemists was 87.5% and the recall of the SDI system was 95.8%. Naturally, the SDI produced about 60% non-relevant citations but the time required to evaluate and reject these was not significant. The truly significant factor is the economic one.

Cost-Effectiveness

The value of an SDI system can be measured in terms of time saved. There are many other values but cost effectiveness is the criteria that is most often applied by the subscriber. Not all cases are so dramatic, but in the example cited above the cost of the manual search using average rates of $20,000/man year was $87,000 whereas the cost for the eleven profiles was $2,800 (or $4,500 using our current price schedule where subscription rates are related to number of profile terms).

An American Chemical Society survey reported in Chemical and Engineering News 47:3;, July 28, 1969 that the average industrial chemist spends 11.8 hours per week in current awareness and literature searching. Of the 7.5 hours spent on current awareness SDI effected an average savings of 3.1 hours for every hour spent. Assuming an expense to the company of $15.00/hours the savings over a year would be almost $2,500. Perhaps much of the time would have been spent in off hours and would not have netted the company additional productive man hours. However, conservatively speaking, a rule of thumb might be that if the user saves only one hour per week, at $15.00/hours the cost is $780.00 which is considerably more than the average cost of $250/year for an SDI profile.

Speed and Regularity

A machine search is done very rapidly and the search for each user is done each week on a regular basis. The machine does not have time off for illness and vacation. The human searcher on the other hand, may become burdened with other tasks and may not have time to go to the library each week. Even in cases where the users area of research is narrow enough and well defined so that he can readily do his own current awareness of CA by searching a limited and manageable number of sections the question remains, will he do it and will he do it on a regular basis? If the library is located some distance from his office or if he is one of many users on a long distribution list, he is unlikely to read the current issue when it is published. SDI provides him his output on a regular basis regardless of these circumstances. At IITRI, we have been providing SDI from CA Condensates on a production basis for two years and we have never missed a weekly run due to anything related to our system. (In one instance the supplier was late in producing the tape causing a delay in the search).

Timeliness

The magnetic tape version of a secondary source is usually made available prior to the publication of the hard copy. At IITRI the Condensates search tape is received 1-2 weeks prior to publication of the Chemical Abstracts. The tape is held for a week to provide a backup copy in case any issue is deleted in production or in the U.S. mail. The search is conducted and output received by users approximately 1-2 days prior to receipt of the hard copy in their library. This enables the user to check abstracts at the time he reviews his output citation cards.

Multiplicity of Data Bases

Many organizations require journal coverage from a variety of sources in widely divergent disciplines. They need to use several data bases or secondary sources and if this is done in-house they must be aware of the coverage, terminology and indexing practices of the different suppliers. And, as changes in the data base occur they must be accommodated by the search in preparing search terms and strategies.

Automatic Preparation of Files in Standardized Format

The output of an SDI search is citations that are printed in a standardized format. In cases, such as our system at IITRI, the citations are printed, together with index terms

and hit terms, on cards thus providing a unit record file from which irrelevant or obsolete material can be deleted. In contrast with printouts on computer paper the citations can be sorted and filed according to the user's preference and it is not necessary to retain irrelevant citations. The user's file can be purged and rearranged as needed.

Even in cases where a user can search the current issues of CA or another data base it might be worth it to subscribe to SDI solely for the convenience of having the appropriate references printed on cards in standard format. Few scientists relish copying citations out of CA nor is it always convenient for a secretary to go to the library to type them.

Cost of Data Base Preparation and Operation of an SDI System vs. Subscriptions

Cost of Data Base Preparation

The cost of preparing a data base, which is borne by the suppliers and incurred only once, is high. The information center and subsequently the subscriber pays only a small fraction of the cost.

The Condensates data base covers approximately 20,000 primary journals, includes more than 25,000 new citations per month, and the cost of preparing this data base in machine readable form is very high. It is unlikely that an individual company or other user organization could afford to select, abstract, edit, index, keypunch, compose, print and otherwise prepare such a collection. Even preparing a data base from the journals known to be of interest to an individual company would probably cost much more than the total subscription fees required to provide SDI from the same sources.

Clearly, data base preparation is an expensive task and user organizations would not want to duplicate the efforts of the data base suppliers. The next alternative is that of processing data bases in-house and operating an SDI system. This too, is more expensive than one might anticipate. The mere purchase or lease of a data base is not all that is required to run an SDI system. In fact, the costs associated with obtaining a data base are extremely small relative to the balance of the expenses.

Cost of Operation of an SDI System

Operating costs can be considered as four in kind: data base; materials, equipment and furniture; machine time; and personnel. Details for some of the specific

cost items are given in Figure 9.

Data Base Related Costs:

- Purchase or lease of data base
- Royalty payment to data base supplier for all hits or citations disseminated

Materials, Equipment and Furniture:

- Purchase or lease of keypunch and terminals if needed
- Purchase of expendable materials e.g. cards, paper products, postage, office supplies and various out-of-pocket expenditures including travel
- Furniture e.g.: special file cabinets to hold tapes, printouts, listings, etc., and standard office furniture.

Machine Time:

- Production:
  - Reformat data base
  - Edit profiles
  - Prepare profile input for search
  - Search
  - Sort and format output
  - Print
- Statistics:
  - Tape library maintenance
  - Research and Development

Personnel Time:

- Management
- Marketing
- Systems design
- Programming
- Profile maintenance
- Keypunch
- Clerical

Figure 9. Center Costs

The efficient operation of a center requires (1) A management component to direct and oversee all production, research and development activities; (2) A marketing component to develop a market and ensure use of the center (3) A systems designer is required where software systems are to be designed in house--this type of staff is not essential where available software is used (4) Programming capabilities are required to: (a) develop software (b) modify existing software for internal use, (c) change and improve software, (d) adapt to changes in operating system, compiler and configuration and (e) adapt to changes imposed on the center by data base suppliers such as new data elements, changes in data base format, changes in record format, changes in machine code, changes in storage densities, etc. (5) A subject specialist is needed to handle profiles and maintain liaison with users.

Profile coordination involves writing profiles; monitoring profile performance, output and user response; keeping abreast of data base changes such as indexing procedures, vocabulary, subject and journal coverages etc., updating and modifying profiles in response to changes in user interest or in data bases; and maintaining user records with respect to precision values, number of hits and any other type of statistic the center may wish to maintain.

The data base expenditures and materials are the smallest items in a center's budget. In our own case they amount to less than five percent of our expenses. Computer time and personnel time are the major cost elements. Some people have been rushed into thinking that for the relatively minimal investment of $5,000 to $10,000 for acquisition of data bases they can process tapes in-house or operate centers. This is a very unrealistic outlook. A rule of thumb might be that efficient processing of a data base would require at least 100-200 users and an information center needs about $100,000 per year to operate. This is a broad generalization, since I fully realize that software efficiences and personnel requirements will vary among centers depending on the overall system computer installation, quality of profiles, requirements of users and services provided. Bear in mind that these costs assume that software is already available and need not be developed.

Cost of Subscription

The expense involved in purchasing subscriptions to

the SDI services of information centers is by far the most economical way of providing SDI to the average company. Individual profiles range in cost from approximately $100 to $500 per year depending on the center, the data base, the number of terms and the volume of output. The average company can purchase many subscriptions to many data bases for a small fraction of the cost required to process in-house. Use of existing centers provides the user organization with far more flexibility in terms of varieties of data bases, types of services, etc. than could be experienced in-house. The user company has no commitment to a hardware configuration or software package, nor does he have a payroll to meet.

The number of centers in operation in the U.S. today is more than adequate to meet the needs of the limited market. They and their sponsors have borne the cost and headaches associated with design and development, and they are now ready to share the fruits of these efforts by providing service to industry, academic institutions and government facilities. For the most part, development costs are not passed on to users in subscription fees -- only operational costs -- so it behooves members of the scientific community to take advantage of the investment.